DOCUMENT RESUME

ED 412 774 FL 801 193

AUTHOR	Lumley, Tom
TITLE	Assessment of Second Language Performance.
INSTITUTION	National Languages and Literacy Inst., Melbourne (Australia).
PUB DATE	1996-06-00
NOTE	6p.
PUB TYPE	Collected Works - Serials (022) -- Reports - Evaluative (142)
JOURNAL CIT	Digest of Australian Languages and Literacy Issues; n11 Jun 1996
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Allied Health Occupations Education; College Students; Educational Trends; Foreign Countries; Higher Education; Language Teachers; *Language Tests; *Performance Tests; Predictive Validity; Second Language Instruction; *Second Languages; Teacher Certification; *Teacher Competency Testing; Tourism; Trend Analysis; *Vocational English (Second Language)
IDENTIFIERS	*Australia

ABSTRACT

A discussion of current second language testing trends and practices in Australia focuses on the use of performance assessment, providing examples of its application in four specific contexts: an occupational English test used for to assess job-related English language skills as part of the certification procedure for health professionals; performance tests developed to assess the language skills of second language teachers; an oral interaction test for tour guides; and English-as-a-Second-Language tests for prospective university students. Issues discussed in these contexts include how tasks are selected for inclusion in the tests, what really gets assessed in a performance test, whether overall language proficiency can be assessed with a performance test, whether such assessment can be fair, whether abilities other than productive performance can be tested, and the advantages and disadvantages of this form of testing. Contains 19 references. (MSE) (Adjunct ERIC Clearinghouse on Literacy Education)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 412 774

Language Australia

The National Languages and Literacy Institute of Australia Limited

THE DIGEST

of Australian languages and literacy issues

JUNE 1996, No. 11

Assessment of Second Language Performance

Tom Lumley, Language Australia-Language Testing Research Centre (LTRC), The University of Melbourne

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

David Dickson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

FL 801193

BEST COPY AVAILABLE

Language Australia

The National Languages and Literacy Institute of Australia Limited

THE DIGEST

of Australian languages and literacy issues

JUNE 1996, No.11

Assessment of Second Language Performance

Tom Lumley, Language Australia-Language Testing Research Centre (LTRC), The University of Melbourne

This *Digest* will explore the notion of second language performance assessment with the aim of explaining some concepts, illustrating them with examples of the ways in which performance assessment is carried out in Australia, and drawing attention to issues that need to be considered in the design and administration of performance assessment.

What is second language performance?

The idea of second language performance usually includes both the ability to *manipulate* the rule systems and formal features of a language (vocabulary, grammar, sentence structure, spelling, pronunciation etc), as well as the ability to *use* language appropriately in a given context. Performance (what you do with the language) is thus distinguished from knowledge of the rules and formal features of the language.

What does performance assessment look like?

Broadly, performance assessment is the measurement of the ability of candidates to *perform* particular types of language tasks. These tasks may relate to general language use, or be relevant to a given context. When determining a person's general proficiency in a language, or when it is considered to be impossible to specify in any detail the sort of contexts in which a person may have to operate, assessment tasks may relate very generally to work situations, or to language use in social situations. However, performance assessment is perhaps more common when the contexts of language use can be more clearly specified. Examples of specific contexts are:

- working as a general medical practitioner
- teaching Italian to primary school children
- acting as a guide for groups of Japanese tourists
- studying at an Australian university.

This paper will focus on this last type of performance assessment in specific contexts such as these. Performance-based assessment involves using evaluative language tasks which relate to what people are required to do in the real world. Because of this real-life focus, they are commonly used for accreditation purposes in professional or academic situations such as those mentioned above.

Three types of performance assessment have been described in the literature (Wesche 1992):

- the observation over time of individuals as they carry out their normal work routine
- the assessment of performance on a number of specified tasks within the actual work setting
- the assessment of performance on simulations of specific occupational tasks.

The third approach, using simulations of tasks which occur in specific occupations, is the most commonly used in the assessment of language skills, as it is the most practical. Of the three types of performance assessment listed above, it is the least complex and time-consuming; does not require that the person being assessed is already working in the target situation (eg, as a practising physiotherapist); and allows for standardisation of tasks, so that all people being assessed are presented with the same tasks thereby making the test fair and easy to compare across candidates. Standardisation is obviously a vital issue when one is dealing with assessments that carry high stakes, such as the accreditation of people to practise in the professions.

Example 1: Health professionals

The Occupational English Test (OET) is used as part of the accreditation procedure for health professionals trained overseas who wish to practise professionally in Australia: dentists, dietitians, doctors, nurses, occupational therapists, pharmacists, physiotherapists, podiatrists, radiographers, speech pathologists and vets. If we need to know, for example, whether or not someone has adequate English proficiency to work as a general practitioner, we would be interested in things such as their ability to write a letter of referral to another doctor, or to conduct a consultation with a patient. Simulations of tasks such as these are developed for each version of the OET by staff at the LTRC. For example, in the speaking section of the OET, candidates are required to take part with an interviewer in role plays simulating patient-health professional consultations, where the interviewer adopts the role of a patient and the test candidate assumes his / her professional role. The role plays require some sort of negotiation between the participants, rather than a simple transfer of information, since this sort of communication reflects the real-life communicative

BEST COPY AVAILABLE

demands of health professionals. The materials are prepared according to careful specifications, in conjunction with representatives from each of the health professions that use the test. In order to provide the fairest possible test for everyone, the same selection of materials is presented to all the members of each profession at any test administration.

How do you decide what tasks to use in performance tests?

The tasks selected for performance assessments need to be carefully chosen. Clearly, it is impossible to present the full range of the types of communication required in occupations involving a wide range of communicative acts (to get an idea of the magnitude of the task, try writing down all the tasks you perform in a week using language as part of your professional or academic work). Representatives of the relevant occupation are therefore generally involved in providing information about those tasks which are most relevant and crucial. This stage of test development is called a needs analysis or job analysis.

Example 2: LOTE teachers

Language Australia-LTRC has designed two performance tests of language proficiency for LOTE teachers, one in Italian (Elder 1994) and one in Japanese, while a third test, for Indonesian, is in the process of development. They have the following purposes:

- to certify LOTE teachers
- to select applicants for LOTE teacher education
- to identify professional development needs.

During the job analysis phase of the Italian test, four Italian teaching programmes in different schools were used as sites for observing foreign language teachers in action. These programmes represented various approaches to the teaching of Italian: partial immersion, activity-based, grammar-based, thematic. This observation phase enabled the test developers to obtain a sense of the range of communicative demands teachers face as part of their job; the frequency of particular kinds of communication; and the importance of being able to handle particular types of communicative situations.

The test developers used this information to produce a trial version of the test, including a wide range of tasks types. A number of these were eliminated, following trials, either because they did not work very well as test tasks, or because they were unpopular with trial candidates or raters, or because no additional information about test takers was obtained by including them. The final version of the test aimed to cover the range of types of interaction a teacher of Italian would face in her / his work, both inside and outside the classroom.

What really gets assessed in a performance test?

It is essential to differentiate between two aspects of performance assessment. The first relates to task fulfilment, or successful completion of the demands of the situation or task presented in the test. A language test for a particular profession will employ a range of assessment tasks based on the types of interactions typically encountered by members of that profession. A doctor could be required to successfully reassure a patient, or make an appropriate medical diagnosis; a tour guide might be asked to deal with a tourist's complaint about a hotel room; a teacher of Italian could be required to explain why a student was in error; a student may be instructed to write a convincing explanation of a scientific phenomenon for a university lecturer.

The other aspect of performance assessment, more important here, is based on linguistic performance, where the aim is to establish whether or not the test taker has sufficient language ability to participate appropriately in the sort of situation simulated in the assessment task. If the test taker does not produce a good performance, it is important for test developers as well as raters (assessors) to consider whether this is the result of a language problem or a problem associated with occupational knowledge, competence or experience.

McNamara (1996) discusses two approaches to performance assessment, 'strong' and 'weak'. In his terms, a 'strong' approach is one where success in the task is crucial to success in the test. This makes it more than a language test: the task is the target as well as the vehicle for assessment, the assessment is concerned with effectiveness of task performance against real-world criteria, and consequently both language and content are assessed. On the other hand, in a 'weak' approach, the one used in most language tests, the assessment task merely provides a context for eliciting a relevant language sample. It simulates, but does not claim to replicate, the real world: the task is the vehicle for eliciting language, and the assessment criteria are concerned only with the quality of this language sample. Language and content may need to be explicitly distinguished in procedures which assess both, as the following example shows.

Example 3: Tour guides

Another performance test developed by the LTRC is the Japanese Tests for Tour Guides (Brown 1994, 1995). This test of oral interaction contains tasks simulating the kinds of situations Japanese-speaking tour guides will face in their work. The raters have experience either as tour guides or as teachers of Japanese, or sometimes as both. In order to distinguish the linguistic aspects of the test candidates' performance from their ability to behave like competent tour guides, assessment criteria are divided into two categories: one set relates to linguistic performance and the other to professional competence. Candidates' test performances are scored by trained raters on both categories, and separate reports for each category are provided in the certificates they receive. This allows raters to separate, in their assessments, the decision about whether or not the test taker would make a good *tour guide* from the issue of whether or not he / she has enough *language* to communicate well in Japanese.

Can we predict overall language proficiency in a performance test?

It is a complex business to talk with certainty about what test takers can do on the basis of small samples of language. Part of the difficulty for any assessment procedure rests in the question of whether it can to be truly representative of the range of communicative ability that it claims to test.

BEST COPY AVAILABLE

Some researchers (eg, Bachman 1990) have argued for the importance of 'construct validity' in test construction and for a more precise analysis of the critical features of communicative language use. According to this view, performance testing becomes the testing not of authentic texts but of the authentic features which underline such texts. The point here is that the act of selecting tasks that appear to simulate the real world is not in itself sufficient. Rather the need is to ensure that tasks allow assessment of the interactional abilities underlying performance on the task — abilities which are transferable to other situations than the task specifically offered during the assessment procedure.

Nevertheless, there are certain tasks that we would want to feel confident that candidates can perform successfully when they are in particular occupations. For example, a doctor must know how to negotiate a course of action with a patient, and a LOTE teacher should be able to give a set of instructions to a group of learners in the target language. While the main interest is in the underlying language abilities (Bachman 1990), Davies (1995) has argued for an integration of this aim with careful sampling from the domain of the profession, to ensure that the underlying language abilities required are adequately represented.

Performance tests are generally rated according to a variety of performance criteria. These may be stated in general terms (eg, fluency, intelligibility, resources of grammar and expression, coherence and cohesion), or they may relate specifically to the tasks included in the assessment (eg, 'Recognises a range of workplace safety signs'; Manidis and Jones 1992). The latter is particularly common in competency-based assessment. The apparent advantage of relating performance criteria to specific tasks (for face validity and ease of assessment) must be weighed against the need to provide assessments which have 'generalisabilty', that is, they can tell us more than just how well the candidate performed on one particular task on one particular occasion.

An associated issue is the need to recognise that language users who attempt a particular language task will show varying levels of performance on the task. It is misleading to assume that a task belongs to a level. Instead it is necessary to recognise that the level of performance will be decided during the assessment carried out by the rater. A task requiring language learners to compose a short formal letter or to comprehend a novel, for example, has the potential to elicit a very wide range of levels of performance; the assessment criteria used and the levels they describe are the most appropriate way of determining a fair assessment of the writing.

Can performance assessment be fair?

Performance assessment relies on subjective judgements carried out by raters who are most commonly language teachers or others with language training of some kind. The training of raters in the scoring of performance is necessary to improve the reliability (dependability) of the assessments made, since without this training significant discrepancies between individual raters are almost inevitable (Lumley and McNamara 1995). This training will reduce but not eliminate differences between raters (Weigle 1994; Shohamy, Gordon and Kraemer 1992). When the assessment procedure involves high stakes (such as entry to a profession or to an educational institution), the unavoidable uncertainty associated with subjective assessments requires a minimum of two raters to be involved, and continual monitoring of rater behaviour (consistency and level of harshness) in order for the person being assessed to have a chance of being fairly treated and a valid assessment to be made (Davidson 1992; Lumley and McNamara 1995). The conditions under which an assessment procedure is administered will also exert a major influence on performance (Bachman 1990; O'Loughlin 1995; Lumley and Brown, forthcoming). It is necessary, therefore, that such variables as texts and tasks used in the assessment, interviewer behaviour and time allowed for candidate response are carefully controlled, if significant weight is to be attached to the assessment.

These measures aim to improve the reliability (or dependability) of the assessment by controlling variables of the assessment procedure. There are other, equally significant aspects of language assessment which also affect its fairness. These include a range of subjective decisions made in the process of test development and administration, which will vary according to who makes them and on what basis they do so (for an example see Alderson, Clapham and Wall 1995). These decisions affect: the test specifications (including content, tasks and items used, and their design); the content of the assessment criteria and/or scales used; the interpretations made of test scores; and the setting of standards (how much is enough for the specified purpose).

Do performance tests assess only productive language ability?

Performance tests are most commonly thought of as tests of productive language: can the learner say or write what is needed for a particular context? However, it is possible for performance tests also to include components which assess the receptive skills of reading or listening comprehension. A writing task, for example, might be based on a reading text that test takers have to read and comprehend before writing their answer. This approach is used in the OET, where the task in the writing test is to produce a letter of referral. Because this part of the test is designed also to assess test takers' reading ability, they are presented with a set of patients' case notes, a type of reading material they can expect to encounter often in their professional career. Unless they have first read and understood these notes, it will be difficult for them to produce an appropriate answer. On the other hand, of course, unless they can write intelligibly, it will be hard for them to show that they have understood the input text. Performance tests which integrate language skills in this way are common.

Example 4: University students

A listening test, too, may include tasks which simulate real-life language use. A number of tests are currently used in Australia to determine the English language proficiency of prospective university students. One of those used in Victoria is the University Test of English as a Second Language (UTESL) (Lumley 1993; Hill and Viete 1994). The listening sub-test of the UTESL takes the form of a short lecture, from which test takers have to take brief notes. In

5 BEST COPY AVAILABLE

this way, the text used simulates not only the type of language students will have to understand at university, but also the task that they will have to perform, ie, taking notes.

What are the advantages and disadvantages of performance assessment?

Performance assessments have been criticised for a number of reasons. Firstly, they are relatively costly to administer, particularly if circumstances require the use of more than a single rater. Secondly, the subjective nature of assessments of this kind mean that reliability may not be very high. This potential problem needs to be weighed against the potential for greater validity offered by performance assessment. Questions may also be raised about the extent to which tasks performed in the artificial test environment can in fact relate to real life.

Performance-based assessment nevertheless offers a number of advantages. This type of assessment usually has greater face validity than some other assessment procedures because of its requirement for candidates to demonstrate their ability actually to use the language (eg, a test in which someone is required to produce a letter, compared to one where candidates are required to recognise the correctly written forms from a range of choices given). This may serve to motivate language learners to produce their best performance.

An indirect advantage is that the content of performance-based assessment may have a beneficial influence (washback) on the curriculum to which it is frequently related (although as Wall and Alderson [1993] point out, whether and how washback affects teaching is still poorly understood). By explicitly linking language learning and language use in the real world, real-life target language use may become more widespread in language teaching classrooms.

Perhaps the greatest advantage of performance assessment is that it aims to measure performance on tasks which require that learning be applied in an actual or simulated setting. A high degree of realism is provided to the test situation by the test stimulus or by the expected response or both. This should result in better predictions of test takers' ability to communicate successfully in real-life situations.

References

Alderson, J.C., C. Clapham and D. Wall. 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Bachman, L.F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Brown, A. 1994. 'LSP Testing: the role of linguistic and real-world criteria.' In Khoo, R. (ed), *LSP: Problems and Prospects*, 202-218. Singapore: SEAMEO Regional Language Centre.

Brown, A. 1995. 'The effect of rate variables in the development of an occupation-specific language performance test.' *Language Testing*, 12, 1: 1-15.

Davidson, F. 1992. 'Statistical support for training in ESL composition rating.' In Hamp-Lyons, L. (ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood, NJ: Ablex.

Davies, A. 1995. 'Testing communicative language or testing language communicatively: what? how?' *Melbourne Papers in Language Testing*, 4,1: 1-20.

Elder, C. 1994. 'Performance testing as benchmark for second language teacher education.' *Babel* 29, 2: 9-19. Also published (1994) in *Melbourne Papers Language Testing* 3,1: 1-26.

Hill, K. and R. Viete 1994. 'Testing English Proficiency for Entry to Diploma in Education Courses.' *Australian Linguistic Matters*, Vol 2, No 3.

Lumley, T. 1993. 'The notion of reading comprehension sub-skills: an EAP example'. *Language Testing*, 10, 3: 211-234.

Lumley, T. and A. Brown. Forthcoming. 'The role of professional informants in content validation of specific-purpose language performance tests.' In Wigglesworth, J. and C. Elder (eds.) *The Testing Cycle: New Perspectives*. ARAL Series S.

Lumley, T. and T.F. McNamara. 1995. 'Rater characteristics and rater bias: implications for training'. *Language Testing*, 12,1: 54-71.

Manidis, M. and M. Jones. 1992. *Certificate in Spoken and Written English. Description of Competencies: Stage 1,2 and 3*. Sydney: NSW Adult Migrant English Service.

McNamara, T.F. 1990. 'Assessing the Second Language Proficiency of Health Professionals'. Unpublished PhD Thesis, the University Of Melbourne.

McNamara, T.F. 1996. *Measuring Second Language Performance*. London: Longman.

O'Loughlin, K. Forthcoming. 'Test taker performance in direct and semi-direct versions of the oral interaction sub-test.' In G. Wigglesworth and Brindley, G. (eds.) *access: Issues in English Language Test Design and Delivery*. Sydney: NCELTR.

Shohamy, E., C. M. Gordon and R. Kraemer. 1992. 'The effect of raters' background and training on the reliability of direct writing tests.' *The Modern Language Journal*, 76, 1: 27-33.

Wall, D. and J.C. Alderson. 1993. 'Examining washback: the Sri Lankan impact study'. *Language Testing*, 10, 1: 41-70.

Weigle, S.C. 1994. 'Effects of training on raters of ESL compositions.' *Language Testing*, 11, 2: 197-223.

Wesche, M.B. 1992. 'Performance testing for work-related second language assessment.' In Shohamy, E. and R. Walton (eds) *Language Assessment for Feedback: Testing and Other Strategies*. Dubuque, Iowa: National Foreign Language Center Publications.

For more information about any of the tests mentioned in this paper, contact Language Australia-LTRC, 149 Barry Street, The University of Melbourne, Parkville, VIC, 3052. Tel. (03) 9344 4389; Fax (03) 9344 5163; email <ltrc_enquiries@muwayf.unimelb.edu.au>.

LANGUAGE AUSTRALIA

The National Languages and Literacy Institute of Australia

Language Australia Limited

PO Box 31, Belconnen ACT 2616 Australia

Tel: (06) 253 4622 Fax: (06) 253 4626 email: NLLIA@nllia.gov.au

ACN 051 438 825. ISSN 1 875578 36 6.

© Language Australia 1996

Readers are encouraged to distribute copies of *The Digest* in its entirety.